Received: 16 August 2018 | Revised: 28 December 2018 | Accepted: 2 January 2019
DOI: 10.1002/widm.1304

WILEY WIREs DATA MINING AND KNOWLEDGE DISCOVERY

OVERVIEW

Computational communication research

Sorin A. Matei[1] | Kerk F. Kee[2]

[1]Brian Lamb School of Communication, Purdue University, West Lafayette, Indiana

[2]College of Media & Communication, Texas Tech University, Lubbock, Texas

Correspondence
Sorin A. Matei, Brian Lamb School of Communication, Purdue University, 100 N. University Street, BRNG 1290P, West Lafayette, IN 47907.
Email: smatei@purdue.edu

Funding information
College of Liberal Arts, Purdue University, Grant/Award Number: 0001

The article advances a new way to think about computational research, in general, and computational communication research, in particular. Contrary to current definitions of "computational science," which emphasizes its inductive nature, we define computational research as an incomplete inductive process, blending both theoretical and data-driven methods of discovery. Communication theory needs to be driven by a clear concept of human needs and abilities, recovering and extending known theoretical insights from mass and interpersonal communication research. The definition we propose for computational communication research has a practical implication. Relying on theory, the definition demands to identify specific processes and domains within the field of computational communication research. The processes include communication production, behavior, and effects. The domains include collaboration, trust, and data storytelling and journalism, while the methods include content and network analyses. The article starts with a broad definition of the "computational" approach, using the Johari window. We continue with a typology of computational communication research, which blends reviews of foundational texts with summaries of leading research. In the conclusions, we discuss the strengths and identify new opportunities in the field of computational communication research.

This article is categorized under:
Fundamental Concepts of Data and Knowledge > Human Centricity and User Interaction
Commercial, Legal, and Ethical Issues > Social Considerations
Fundamental Concepts of Data and Knowledge > Big Data Mining

KEYWORDS
communication, computational science, research methods

1 | INTRODUCTION: WHAT IS COMPUTATIONAL COMMUNICATION RESEARCH?

Computational science, of which a good part of computing-intensive communication research is a branch, is a new way to think about and overcome some of the limits of human inquiry. Its main claim is ambitious: given enough data and computing power, all discovery problems become tractable. Started in biology, where it led to the sequencing of the human genome, computational science has been increasingly adopted in social sciences. Computational social science was proposed and defined as a field of inquiry in a *Science* editorial by a team of social and natural scientists working in the emergent domain of "network science" (Lazer et al., 2009). Yet, the definition they proposed for a future "computational social science" was elusive. Instead of using the classical "closest genre and specific difference," the authors defined computational social science by describing its means and consequences. According to them, computational social science, is a type of inquiry "that leverages the capacity to collect and analyze data with an unprecedented breadth and depth and scale [to explore] how people interact [and to] offer qualitatively new perspectives on collective human behavior" (Lazer et al., 2009, p. 722). This definition leaves out a crucial question: how is computational science essentially different from noncomputational science? Are the two the same in all respects except for scale? Are there any differences regarding units of analysis, domains of most appropriate inquiry, application realms, and so on? What types of question can and what types cannot be asked through a computational approach?

The questions remain relevant when we move from social to communication science, which is the object of this article. In fact, they become even more complex, since communication investigates multiple domains, is interested in realities at various levels of analysis and complexity, and demands very practical answers in many cases.

In what follows, we will propose a definition of the computational approach and a breakdown of its domains of investigation and possible outcomes. We will also propose some possible applications and areas of impact. Throughout, we will also propose to combine inductive with deductive approaches, which leverages theory, without privileging it.

Any survey of computational communication research should start with defining the boundaries of the term. Such boundaries can be set by answering two simple, yet foundational questions: What is the meaning of the adjective "computational"? How is computational communication research different from "simple" or noncomputational/traditional communication research? These questions are justified by the fact that rare is the researcher in communication who does not use one or another form of computing to collect, store, sort, and analyze the facts his or her work pertains to. The digital nature of the scientific work aside, one other way to answer the questions is by saying that the adjective "computational" refers to the application of vast, enterprise-level computerized resources to analyze big data that refer to large, and by size, significant issues.

However, should size matter? Science is science. If the workflow is the same, any scholarly enterprise should be like any other, regardless of the amount of data involved. Social and communication science should be assessed by how generalizable their conclusions are, not by how large their samples and datasets are. If the conclusions drawn on large samples only add a bit to the significance levels of the results, "computational" approaches could be considered a matter of technology, technics even, not of major intellectual significance. Thus, the adjective "computational" might not mean or add much to the conversation unless it carries new epistemological attributes.

In the next section, which opens the discussion about computational communication research, we will focus on two essential attributes and their necessary reinterpretation in the context of communication research: an expected emphasis on inductive discovery and the primacy of population-size, networked phenomena.

1.1 | Computational science: Inductive, deductive, or both?

Computational research, be it in social, physical, or life sciences, should not be equated to digital or even to "big data" research. We need to go to the original meaning attached to the computational method. According to its original proponents in bioinformatics (Hagen, 2000), the computational approach is and should be different from other types of discovery enterprises by the unique way of marshaling the data and by the analyses it makes possible to provide new and unexpected results. According to Watts (January 2017), discovery by computational means should overcome the limits imposed by existing theory. Computational approaches should focus on practical, that is, empirical problems for which we might not even have a theory.

What Watts suggests, although not in these terms, is that computational science should be mainly preoccupied with "unknown-unknowns." In reformulating the problem this way, we use the "Johari window" originally proposed in psychology (Luft & Ingham, 1955) and often used in project management (Girard & Girard, 2009, p. 54). The window proposes a four-quadrant division of knowledge. Facts that we know or not are intersected with the awareness of what can be known with present methods and theories.

	WE ARE AWARE	WE ARE NOT AWARE
KNOWN FACTS	KNOWN-KNOWNS	UNKNOWN-KNOWNS
UNKNOWN FACTS	KNOWN-UNKNOWNS	UNKNOWN-UNKNOWNS

When applied to epistemology, the window makes visible the elementary idea that the human mind in its quest of knowledge discovery mainly handles four types of knowledge: known-knowns, unknown-knowns, known-unknowns, and unknown-unknowns. The known-knowns refer to verified and validated truths: the earth is round. The known-unknowns are the things we know we do not know yet: the exact moment when the sun will die. The unknown knowns are the things that become relevant only in hindsight: the July 2001 "Phoenix" FBI memo that forewarned about a possible attack on U.S. targets with civilian planes (Behar, 2011). Finally, there are the many unknown-unknowns, the many things we do not yet know to be possible or even probable and for which, by definition, we cannot even provide examples other than in retrospect.

The window was made famous by a political debate during the Iraqi war (Rumsfeld, 2002), which although ridiculed as a glib tongue twister ("Foot in Mouth" Prize for Rumsfeld, 2003), it highlighted the complexity of political decision-making. In the preceding months to the Iraqi war, when confronted by a journalist that there was no evidence for a connection between Saddam Hussein and terrorists armed with weapons of mass destruction, Donald Rumsfeld, the secretary of defense, answered in a rather convoluted way. He said that there is no way to know if that is true or not since it is not in the habit of terrorists to flaunt their patrons or of the dictators to openly recognize their secret allies. To make this claim, not only expedient but to add a note of logical necessity, he classified what the world knew about Saddam Hussein's arsenal as of 2013 as "unknown unknown." The way Rumsfeld used the Johari window demands that the famous exchange should be quoted in full:

> **Q (reporter):** In regard to Iraq weapons of mass destruction and terrorists [...] there are reports that there is no evidence of a direct link between Baghdad and some of these terrorist organizations.
>
> **Rumsfeld:** Reports that say that something hasn't happened are always interesting to me, because as we know, there are known knowns; there are things we know we know. We also know there are known unknowns; that is to say we know there are some things we do not know. But there are also unknown unknowns—the ones we don't know we don't know. People who have the omniscience that they can say with high certainty that something has not happened or is not being tried [...] can do things I can't do (Rumsfeld, 2002).

In brief, the classification of knowledge according to the Johari window and re-elaborated during the Rumsfeld debate starts from the premise that before initiating the inquiry process, all knowledge seekers need to take an inventory of what they do or do not know and of what they would like to know. Rumsfeld, rightly or wrongly, put the Saddam Hussein-terrorists in the category of "unknown-unknowns" and pushed it aside.

In science, however, what he pushed aside is the holy grail of discovery. The unknown-unknowns are what fueled some of the most momentous scientific discoveries. Their heuristic value was rightly highlighted by methodologists (Logan, 2009), who similarly believe in the value of unknown-unknowns. To make this idea more tangible, let us remember that it was Fleming's lack of knowledge that penicillin cultures can kill bacteria (a true unknown-unknown in 1928) against which his accidental discovery of the lethal effect of the fungus on certain pathogens was projected, turning his blind spot into a highly relevant and worth pursuing discovery (Abraham, 1980).

The great attractiveness of the computational approach is that unexpected discovery of patterns and connections between data points may open the vault of unknown-unknowns. Mere science of precomputational vintage would or should only deal with known-unknowns, working its way by deductive inference from known to possible theories that are to be validated by empirical research.

Computational science, on the other hand, is seen as an inductive method of answering questions we do not even know how to ask. This is possible due to the fast and vast abilities of computerized methods to churn and sieve very large amounts of data and to find regularities and correspondences that are not immediately apparent to the researchers. They are latent in the data and are revealed algorithmically through network (Hansen, Schneiderman, & Smith, 2010; Smith, Shneiderman, & Himelboim, 2014; Welser et al., 2011; Welser, Gleave, Fisher, & Smith, 2007), cluster, or classifier analysis (Hagen, 2000). Computational discovery is a pursuit akin to "signal" or "pattern recognition" in vast masses of information collected as simple data points. The core problem of this approach is recognizing the signals formally, even before knowing what they mean. Later, from their shapes, sequences, and concatenation, we can advance propositions about their meanings or causes.

This, eminently inductive, method is the one that launched bioinformatics, the first form of computational science (Hagen, 2000). Work on identifying the discrete structure and function of structures in proteins or genetic material relies on massive operations of identifying repeating sequences, their correspondence across samples, their clustering, and finally their function by correlations with known effects (Nesvizhskii, 2014). The process very often involves raw pattern recognition and later hierarchical clustering or principal component analysis. It is the triumph of looking for the proverbial needle in the haystack with the difference that the needle is made of vegetable fiber itself and there are multiple needles of multiple sizes.

Straight borrowing of this type of exploratory analysis from bioinformatics to social science raises several questions. Although there is merit in engaging in more open-ended, practical research, as suggested by Watts (January 2017), it is hard to discard theory, even partially. While in genetics and biology we are just beginning to understand the mechanisms by which genes influence process and characteristics and inductive approaches are still useful (Nesvizhskii, 2014), in social sciences, we have had many and successful attempts to define and model human behaviors. A plethora of social and psychocognitive theories explain how and why individual make choices, organize, or connect to each other.

In borrowing computation approaches, we should not discard theory. Otherwise put, when borrowing the computational approach, we cannot indiscriminately adopt the purely signal recognition or inductive method (Borge-Holthoefer, Moreno, &

Yasseri, 2016). Rather, we need to combine induction with deduction, following the famous recommendation of Francis Bacon (2000) in *The New Organon* to advance toward true knowledge through semi-induction, which is empirical detection of patterns steered and mitigated by theoretical assumptions.

Computational approaches to communication, in fact, cannot and will never be purely inductive. Even when they are presented as such, they carry unspoken assumptions, which are proto-theoretical and rooted in common sense. This can create many problems of internal and external validity. Common sense is notorious for its overreliance on apparent formal similarities and regularities, which often can and should be explained by theoretical principles that are not directly intuitive or apparent. For example, the empirical evidence provided by the senses suggests that the sun and many other stars circle the earth. Yet, contradictory information, such as the apparent retrograde motion of the planets, when they stop moving in one direction and start heading in the opposite one, needs to be integrated by an alternative explanation. This demands a theoretical model that contradicts what the senses tell us, namely, that it is not the earth that is the center of the solar system, but the sun.

1.2 | Population size research, networks, and contagion

The second promise of computational research, after its ability to detect signals in noise, is its ability to handle population-size data. In communication, we can harvest and analyze the entire corpus of edits made on Wikipedia (Matei & Britt, 2017) or on vast collections of wikis of many kinds (Kittur, Lee, & Kraut, 2009; Shaw & Hill, 2014). We can analyze entire storms of interactions created by viral tweets on Twitter. We can mine all the posts and comments committed to an open Facebook group or to a sub-Reddit (a section of the famous electronic bulletin board Reddit.com). We can analyze all the code additions and changes made to the Linux kernel (https://perma.cc/5NNE-7RYR), the most famous open-source software project, and the questions and answers committed to StackOverflow.com, a vast, just-in-time knowledge production system for amateur and advanced programmers alike. For all these datasets, we can get data about the entire population and we can obtain or infer the connections between the members of the population.

While population-size data are interesting, given the logic and efficiency of sampling, its mere size should not impress us that much. In fact, if all we got from computational science were samples as large as the entire population, the cost of obtaining and handling this type of data would not be worth the benefits. Stochastic studies based on samples of the right size are more cost efficient and just as revealing as having access to the entire population. There is one type of data that truly benefits from population-size samples. This is graph data, which records both cases and the relationship between them. The gift of obtaining access to the entire set of relationships between the members of the population, in other words, to the implicit and explicit graphs of connections, is a true boon to research (Hansen et al., 2010; Newman, 2010; Watts, 2003). Typical scientific research, especially in social and communication sciences, relied on the old probabilistic ideal of samples made of randomly selected, independent cases. Such samples eliminated or at least minimized the relationships between the subjects or units of analysis. The goal was to reduce unseen correlates due to case-to-case contagion through communication or proximity.

However, many if not most social phenomena have an autocorrelative component. Communication research should and needs to consider the fact that the individuals or cases studied are connected via implicit or explicit communication. Autocorrelations due to spatial or temporal proximity are or should be focal points of research. Such research, however, cannot be conducted on random samples, but only on intact populations. Members of samples will rarely if ever connect to each other. Many of the connections that influence the members in a random sample are in fact unknown. When we have a full population, however, we have access to all the connections between all members, which with appropriate research methodologies can be used to great effect. We can predict why certain classes of subjects behaved in a certain way not only because of their intrinsic characteristics but because of their connections to each other. Autocorrelation and error term correlation, the vices of purely probabilistic, random, independent samples, become the virtues of population size, computational research.

In a way, we can say that if there is a true virtue in computational research, that is its ability to explore population size problems by taking advantage of the graph approach that such problems make possible. Of course, given that the number of links in a population increases quasi-exponentially with the number of nodes, and even if mitigated by the formula $n*log(n)$, as recommended by Briscoe, Odlyzko, and Tilly (2006), computational research of predictive graphs could be quite costly, both in terms of computing effort and technical abilities. Most off-the-shelf computational solutions for graph problems stop being feasible above a few thousand nodes. Merely mapping graphs of interaction for gigantic networks, such as the editorial network of interactions for Wikipedia, is often difficult to handle, although solutions and even datasets have been produced (Leskovec & Sosič, 2016). Yet, as we will see below, much promise exists in this field, which might make computational research meet many of its promises of revolutionizing social and communication sciences.

2 | COMPUTATIONAL COMMUNICATION RESEARCH: PROCESSES, DOMAINS, AND METHODS

In what follows, we will survey computational communication research broadly, in view of some foundational works published in the last years. We focus on processes, domains, and methods. Throughout, we will highlight the tension between inductive and deductive approaches, pointing to the constant need and successful use of theory and semi-induction. We chose processes, domains, and methods to highlight the fact that communication research is multidimensional. Our typology involves domains situated at various levels of analysis (group or societal) or that deal with crucial aspects of human interaction, including trust and data storytelling. It covers heterogeneous processes (production, interpersonal and mass–personal communication behavior, or effects of communicative interaction). Of the many possible methods of research, we chose two that promised and delivered the most: text (content) and network analysis. Table 1 maps out our approach. In view of our interest in essential processes that define computational communication analysis, we focus on production, behavior (defined as interaction with content) and the effects of communication processes. In this respect, we fall back on the communication research tradition, which has grown out of questions related to "who?," "how?," and to "what effect?" (Lasswell, 1968). By this, we also want to highlight that computational communication research is a part of and not a secessionist movement from the great tradition of communication inquiry.

The domains were selected for their presence in the communication literature and potential for impact. We spent a good amount of time on collaboration and organization. Communication networks have created new ways to organize and produce content, movements, or businesses. Social media and crowdsourced platforms, such as Uber or Wikipedia, literally changed that face of the world. Studying the processes sustained by these platforms using computational methods helps us see more clearly how the world is changing. Finally, we focus, selectively, on two emerging methods of great importance, content analysis, and predictive network analysis, which have and will provide a richer description of the changes mentioned above. While this classification may change in the future, we propose it here as a simple and effective heuristic method that allows breaking down the problem into smaller, more tractable parts.

2.1 | Foundational texts, questions, and open agenda items

Computational research has received serious attention within the discipline of communication in recent years. This can be seen in the establishing of the Computational Methods Interest Group by the International Communication Association in 2017, and the publication of some foundational articles (e.g., Cappella, 2017; Freelon, 2014; García-Peñalvo, 2016; Shah, Cappella, & Neuman, 2015) on computational research in the field. Cappella (2017) observed that there has been much work outside of communication that takes on questions at the core of the communication discipline. In fact, Shah et al. (2015) argued that the strong trend in computational social science to treat "text-as-data" signaled the rise of computational communication science as a robust emerging field.

The computational approach also requires "computational thinking," an idea advanced by Wing (2006). Computational thinking requires information technology literacy combined with reading, writing, and math competency. García-Peñalvo (2016) makes another contribution to computational thinking, defining it as "the underlying problem-solving cognitive process

TABLE 1 Key processes, domains, and methods of computational communication research

Processes	Production analysis	Social capital, uses and gratifications, sociostructural factors, coordination and conflict, evolutionary approach, social learning, network centrality, media dependency system, media diversity, online media production, and economic mechanisms.
	Consumption/behavioral analysis	Diffusion, contagion, information cascades; behavior, cognitive processes; information consumption, and network embeddedness.
	Effects	Cultural, moral, and sociopsychological effects of social media; concerns for social disconnection, alienation, and depression; nonverbal effects on persuasion, and framing effects on presidential debates
Domains	Social organization and collaboration	Social-structural approach, adhocratic emergence, transactive memory, public goods, structuration, self-organizing, organizational change (life cycle, teleology, dialectics, and evolution), homophily, assortativity, and group productivity/outcomes.
	Trust, credibility and, transparency	Prosumers (content producers and consumers), online reviews, expectancy violation, personal reputation veracity, and endorsement evidence.
	Data journalism	Computer-assisted journalism, programmer–journalists, data visualizations of polarization, textual analysis of tweets, computational propaganda, and algorithmic agents.
Methods	Content analysis	Text-as-data, automated content analysis, challenge of natural language processing, rhetorical analysis, speech act theory, ontological semantics, sociolinguistics, intention analysis, use computational approaches to verify findings based on other approaches and test existing theories.
	Predictive network analysis	Taxonomies of network topologies, network structures are used to predict node characteristics or node attributes are used to predict the existence or emergence of network edge (ties), exponential random graph modeling, network autocorrelation procedures

that allows [coding and programming]" (p. vi). Computational thinking is the ability to apply algorithms and high-level abstraction to solve problems. He argues that while coding is one of the best ways to promote computational thinking, computational thinking is applicable to problems that do not involve coding, including problems outside of computer science.

The main asset of computational communication research is the opportunity to collect and analyze mass quantities of digital data, such as trace data (Crowston, 2017). However, the computer programming skills necessary for this task are sparsely distributed among communication researchers, as it is not a traditional research method. This is not an unmitigated benefit. Some communication researchers had to stumble along and figure out their own solutions (Freelon, 2014), while others collaborate with researchers in computer science, engineering, statistics/mathematics, information systems, social computing, and high throughput computing (Cappella, 2017; Shah et al., 2015). This, however, can lead to new, broader, and innovative types of data collection and interpretation, which are not bound by tradition (Matei, Russell, & Bertino, 2015). This follows trends that go beyond communication. The computational approach has also been integrated into other social science fields, such as political science, sociology, psychology, public health, and public policy (Shah et al., 2015).

Computational communication research focuses on "natural occurring" actions and interactions, such as online behaviors on Amazon, social expressions on Twitter, and personal connections on Facebook. Lazer et al. (2009) termed this collection of data points "digital breadcrumbs." Shah et al. (2015) maintain that these "digital breadcrumbs" provide a range of influence processes and social phenomena for computational research, such as political behaviors and personality traits. In fact, they argue that it is the emphasis on "natural occurring" actions and interactions that give computational communication research the unique advantage to overcome some of the persistent limitations of traditional research, such as experimenter effects, self-report, and social desirability biases in survey and experimental studies. We argue that such natural data can also help improve existing theories and perhaps challenge weak theories of the past, provided that deductive and inductive methods can be used concurrently.

While the unique advantage of the computational approach is powerful, communication scholars see several potential pitfalls. Shah et al. (2015) point out three distinct challenges: generalizability, ethical concerns, and theoretical scope. The challenge of generalizability refers to the limitation of online data, which reflects demographic profiles that diverge from those of the general population. For example, while Twitter can provide a large amount of data for researchers to identify communication patterns, Twitter users are younger and more educated, not representative of the population (Smith & Anderson, 2018). Furthermore, Lewis, Zamith, and Hermida (2013) report that tweets gathered online can yield incomplete data sets that researchers assume to be complete when they are not. Some events can generate millions of tweets, which cannot be captured with common data harvesting tools. The data obtained are the data that happened to be available at the time when the researcher cast his or her net. Furthermore, as with all big data sets, most statistical analyses will likely turn significant, due to the sheer amount of data. In other words, the p-value is no longer as important and relevant as an indicator of validity in big data (van Atteveldt & Peng, 2018). This can call into question the generalizability of computational research using big data.

Second, the computational use of communication data is fraught with ethical dangers. While data might contain much insight, it is at the same time linked to specific users, who may not know and/or want that their information is being used for research. The lack of formal consent might compound the problem, especially when data was not publicly available. Furthermore, the researchers and the users may not even know, at the time of data collection, how the data may be used or reused in the future, which makes formal consent for a future analysis practically impossible. The issues of data privacy and de-identification become of paramount importance.

Third, and directly relevant to our main thesis, computational analyses of large amounts of data (such as tweets and connections among social network users) can become overly inductive, rendering theory as a secondary concern, or sometimes, even irrelevant Anderson (2008). This is problematic because social data analyses uninformed or underinformed by social theories can be misinterpreted, thus making predictions, forecasts, and explanations that misrepresent reality. Freelon (2014) points out that this can challenge the interdisciplinary space of computational communication research as researchers become more concerned with algorithmic prediction and software development than with theory development. Therefore, Shah et al. (2015) argue for the need to use good theory to make sense of research, "to offer clear a priori predictions and sensible explanations of what are otherwise uninterpretable statistical tests" (p. 12).

Recently, Cappella (2017, p. 546) summarized as an open agenda the four key questions to be pursued in computational communication research. These questions include: (a) "What structures of communicative networks enhance and retard the diffusion of social influence?" (b) "Can we model media effects incorporating both interpersonal connections and mass communication effects?" (c) "How can efficiencies of mass communication campaigns begin to match the effectiveness of messages tailored to the person-as-target?", and (d) "How does the content of interpersonal communication change as it moves from personal exchanges to mediated messages that scale to broader audiences?" Cappella argued that these questions are foundational because they tackle some of the lasting issues at the intersection of mass and interpersonal communication research. At their core, these questions circumscribe the classic issue of social influence in communication research through the diffusion of information, ideas, and memes in the social world.

2.2 | A forward and summary process-based scan of computational communication research

So far, we have reviewed some existing perspectives on computational communication research, which point to the necessity of proposing a unified framework for considering the expectations and possible deliverables surrounding computational communication research. The framework should balance inductive with deductive methods. It should also consider issues such as generalizability, ethics, theory, effects, diffusion, and cross-domain research. To what degree does existing research meet these areas of concern? What is left out? What more can be done? In what follows, we will review some of the current literature, selectively focusing on three domains, which will intersect with the framework proposed above. First, we will look at how the literature deals with the three segments of the communication process: production, consumption/behavior, and effects. Second, we will scan some representative emerging domains of computation communication research, including organization and collaboration, trust, credibility, and transparency, as well as data journalism. Third, we will discuss content analysis and predictive network analysis as the two most prominent methods used in computational communication research. Finally, we propose some further avenues for further research, which might fill in some research gaps.

2.3 | Modalities of investigation by parts of the communication process

Computational communication research is a growing field of research, covering all sides of the communication process, including production, consumption/behaviors, and effects. In this section, we will highlight some of the representative pieces of research in each of these domains, highlighting where is needed, the accomplishments, the promises, and the gaps.

2.3.1 | Production analysis

The core question of this section is "how does computational research explain user contributions to social media"? The answers are many and varied. As early as 20 years ago, Wasko and Faraj (2000) proposed a set of structural factors, including building and maintaining social capital, which can be leveraged for other activities through generalized reciprocity. More recently, using a uses and gratifications framework, Shao, Ross, and Grace (2015) suggested that social media rests on self-expressive and interactional needs. Furthermore, looking at knowledge production sites, Kittur et al. (2009) or Matei, Jabal, and Bertino (2017) advanced a set of sociostructural factors that may structure and motivate social media production. Kittur et al. (2009) focused at coordination and conflict, while Matei and Britt (2017) proposed an evolutionary approach, supported by functional roles and "adhocratic" order. The latter refers to a flexible, user-defined, practical order ran by those who do the most on a social media site. Both studies used measures of inequality, especially entropy, which can capture the degree to which social groups are structured or not.

Empirically, however, the literature is not as rich as it should be. Research on production motivation remains in the realm of theoretical models (Shao et al., 2015). Some good empirical research does blaze new trails (Crowston, Jullien, & Ortega, 2011, 2013). Facebook or Twitter in-house researchers focused on information cascades (Dow, Adamic, & Friggeri, 2013), influence (Ball-Rokeach, 1985), or similar topics. The 10 years old study by Burke, Marlow, and Lento (2009), which claims that social learning is a good predictor for continuing engagement with social networks of newcomers, is also a notable contribution to this effort.

Some contributions on motivations and recruitment of new members on Twitter or older social media are found in the academic literature (Gonzalez-Bailon, Wang, Rivero, Borge-Holthoefer, & Moreno, 2014), which detail the importance of networks and node centrality in attracting new members on Twitter. Lampe and Johnston (2005) focused on the effect of feedback and support on maintaining and propelling contributions to online discussion groups (at the time, Slashdot). Joyce and Kraut (2006) considered the mechanisms by which interaction and communication sustained in online discussion groups.

The research on production can be significantly enhanced with more comprehensive research agendas. Theories such as Media Dependency System (Ball-Rokeach, 1985), which propose that publics and their media depend on each other to satisfy individual and organizational needs should be brought to the limelight. More important, mature economic models of media production can be woven through the fabric of computation communication research. Incidentally, or not, researchers in France, who traditionally are more welcoming to interdisciplinary research, are blazing new trails in this direction. The Laboratory on Cultural and Artistic Industries at University of Paris leads the way (Marty, Rebillard, Pouchot, & Lafouge, 2012; Rebillard & Smyrnaios, 2010). Their work on media diversity, online media production, and economic mechanisms that explain social media work are a welcome addition to the field. At the same time, more needs to be done to utilize mature and complex media-specific theories for explaining social media production at institutional level.

2.3.2 | Behavior analysis

The most vibrant and interesting research in this area is that on influencers and information cascades (Bakshy, Rosenn, Marlow, & Adamic, 2012). Studies highlighted the importance of supernodes, but also of triggering events in releasing information cascades (Dow et al., 2013). The view that everyday users also matter by the sheer volume of content they put out is similarly important. However, we are far from a unified theory of when and where information cascades are likely to occur. In fact, the great enthusiasm for the information cascades might be an albatross, as each cascade might have its own motivation, hard to align with all others under a catch-all theory. Here, again, we are caught in the trap of inductive research, unled by a clear theoretical justification. While information cascades are current and easy to observe, their mere constant occurrence is not a guarantee that they all belong to the same theoretical mechanism. Like the classic theory of diffusion of innovations, this is more of a class of phenomena in search of many different explanatory mechanisms.

Behavioral research on social media with computational means goes beyond diffusion, contagion, and information cascades. There are significant attempts to triangulate behavior, cognitive processes, information consumption, and network embeddedness. O'Donnell and Falk (2015) reviewed an extensive literature looking at the mechanisms by which micro- and macro-behavioral processes interact. They combined social network with neural data to predict the success of persuasive messages. They also pioneered a method of investigation that looks at neuroimages and their association with sentiments as expressed in language. They found a link between the neural mechanisms related to idea evaluation and the linguistic process by which it is retransmitted.

Pont, Mora, Casado, and Mendez (2016) employed mathematical modeling of belief systems to understand how human beings process incoming information to trigger emerging behaviors, especially in social communication. The researchers custom-built software to infer the pattern of emerging behaviors in communication. The model described the information modification and information propagation in complex systems. These studies demonstrate that human behaviors can be computationally studied by triangulating different big and naturally occurring data (e.g., neural data) with traditional data sets (e.g., self-reported network data).

2.3.3 | Effect analysis

Media effects research is one of the most distinguished strands of scholarly endeavor in communication. Concerns about the cultural, moral, or psychological effects of media are as old as the oldest forms of electronic media (Blumer, 1933; Katz, 2001; Merton, Fiske, & Curtis, 1946). Computational methods have been used in several landmark studies, including the infamous research on the effect of the emotional tone of Facebook messages on user's own production of emotional material, which lead to a heated conversation about the ethics of manipulating users without consent (Kramer, Guillory, & Hancock, 2014). The constant concerns are about social disconnection, alienation, or depression (De Choudhury, Counts, & Horvitz, 2013).

In an experiment of virtual peer influence, Sherman, Payton, Hernandez, Greenfield, and Dapretto (2016) found that the brain areas associated with imitation, social cognition, attention, and reward processing were activated when teenage participants saw photos endorsed by peers, including photos of risky behaviors such as drinking and smoking. These teenagers also were more likely to click the "Like" button to endorse the photo, if it received many "Likes," although the endorsements were by strangers they did not know. In another study, Churches, Nicholls, Thiessen, Kohler, and Keage (2014) reported that the brain areas responsible for facial recognition lit up when participants in an experiment were presented with an emoticon of the smiley face [e.g., :-)]. They explained that this emoticon effect is a neural response created solely based on cultural conditioning, as a result of the widespread adoption of computers. The implication of this study is critical. When the social media message received is that of disapproval (instead of a "Like" endorsement) or a sad face [e.g., :-(], it can have a real effect on the recipients, including experiences of social disconnection, alienation, or depression.

De Choudhury et al. (2013) have looked at the communicative indicators of depression in social media messages (Twitter), obtaining a 77% recognition rate for updates posted by clinically depressed individuals. The model operationalizes and connects social, emotional, and behavioral signals expressed on Twitter to generate a social media depression index (SMDI) that can be used to detect depression in populations. More specifically, they studied participants recruited from Amazon Mechanical Turk and their 69K tweets over a 3-month period, looking for content indicative of depression. De Choudhury et al. tested the relationship between depression (operationalized with the help of the Center for Epidemiological Studies Depression Scale) and the language used in their Twitter posts, finding good linguistic markers for depression. The analysis used the Linguistic Inquiry and Word Count, LIWC, http://www.liwc.net/, demonstrating the applicability of existing corpora and linguistic tools to complex communication research.

Based on the resultant statistical models (and support vector machine classifier), the researchers defined a metric referred to as the SMDI, which is a measure of the degree of depression based on the analysis of a large corpus of Twitter posts by users daily. Findings, in fact, align with the U.S. population statistics of depression reported by the Center for Disease Control and Prevention (CDC). This approach is useful in that it can be done more easily and regularly to monitor depression patterns than traditional surveys, such as the Behavioral Risk Surveillance Survey and those by the CDC. It also illustrates a good blending of theoretical and inductive approaches.

Conway and O'Connor (2016) reviewed several dozen studies looking at mental health and social media use, most of them conducted from a computational perspective. It is noticeable that a majority of them are descriptive and measurement focused, rather than inferential highlight the need of explicit theoretical concerns. Despite the claims that social media-induced depression, advanced by researchers who looked at the problem through an experimental or survey lens (Zagorski, 2017), computational methods are yet to embrace theoretically driven causal approaches. This, again, highlights the tension between data-driven versus theoretical work in computational research. Often, empirically derived measures for phenomena such as "depression" are proposed and operationalized before a strong theory of psychological engagement with media has been constructed.

At the same time, the "effects" research of traditional media has built a long and distinguished tradition, advancing from simple behavioral models to complex, cultural–cognitive frameworks, some of which propose a continuous give and take between media consumers and their sources. Computational communication research needs to go back to this tradition and to recover all that could theoretically be recovered. Foundational texts, such as Sparks (2015), should be constantly referred to in the crosshair of the research community. Furthermore, the compendia of current and recent models of media effects research, infused by deep and well-considered cognitive and psychological theories, including foundational texts (Bryant, Thompson, & Finklea, 2012) need to be brought to bear in any conversation about the sociopsychological effects of social media.

Fortunately, a limited sample of such theory infused research does exist, although not on the necessary scale. For example, a study of video reviews of movies (Park, Shim, Chatterjee, Sagae, & Morency, 2014) tested the effect of multimodal computational descriptors of verbal and nonverbal behavior against unimodal descriptors in predicting a speaker's persuasiveness. They used the Coverep software to collect and quantify speakers' acoustic (such as voice qualities), verbal and paraverbal descriptors (such as articulation rate, pause, pause filler, speech disturbance ratio, and stutter), and visual descriptors (such as facial expressions). They used the descriptors as independent variables predicting perceived persuasiveness of speakers. Their study provided an example of how video data can be computationally analyzed for their persuasive effects on viewers using a mature theory of persuasiveness.

Similarly, Shah et al. (2015) measured tweets in response to a live presidential debate, comparing the volume and sentiment of real-time tweets during the Obama–Romney debate in October 2012. The theoretical lens was framing. They found that the highest tweet volume moment for Obama was when he made a comment that became a meme on Twitter. Their study pioneered the computational approach of linking data related to the "first screen" (televised presidential debate) and the "second screen" (social media expressions on Twitter).

2.4 | Domains of computational communication research

Computational communication can and does cover a large gamut of domains and interests. In this overview, we cannot cover them all. We will only look at some domains, which act as trailblazers both by intrinsic importance and by the attention they attracted. We will focus mostly on issues of social organization and collaboration, trust, credibility, and transparency, and data journalism. For each, we will highlight the crucial points explored by the current research and the gaps that still exist and need to be closed.

2.4.1 | Social organization and collaboration research

Social media has brought with it new forms of organization and collaboration. One can say that it in effect reinvented collaboration and organization. Facebook, Twitter, Wikipedia, Flickr, and YouTube are gigantic content factories and laboratories (Matei, Jullien, & Goggins, 2017). Individuals work with each other in a variety of ways, sometimes openly, other times implicitly (Crowston, 2017). Leaders sometimes emerge and other times are hard to pinpoint. Overall, the literature converged on a strong social-structural approach, which emphasizes the need to understand social media as a space of cocreation via adhocratic roles and emergent informal social structures (Birkinshaw & Ridderstråle, 2010; Dolan, 2010; Konieczny, 2010).

This view replaced the earlier expectations of emergence and wise-crowd phenomena, by which social media was vested with the ability to build new forms of collaboration in the absence of social coordination and control (Brafman & Beckstrom, 2006; Johnson, 2001; Surowiecki, 2004). Computational research emphasized the importance of a small group of contributors (Crowston et al., 2011; Kittur et al., 2013; Matei & Britt, 2017; Ortega, Gonzalez-Barahona, & Robles, 2008), who spend a long period of time on the project and structure the content and the process of collaboration itself. Matei and Britt (2017) propose as an explanatory framework "adhocratic emergence." They identify discrete phases in the emergence of all online collaborative groups as well as a group of "functional leaders," who control the editorial process by their incessant presence in the project and by leading from the front. The mass of their contributions (the top 1% contributors to Wikipedia are responsible for almost 80% of the site's content) is vital for the success of any project. Furthermore, projects that fail in garnering the time and energy investment of such leaders cannot hope to survive in the long term.

The structuration of online groups and teams is not, of course, a new phenomenon. The research tradition of transactive memory and public goods theories summarized by Hollingshead, Fulk, and Monge's (2002) stand behind it. Groups that organize member expertise using online platforms enjoy benefits derived from a more decentralized yet denser communication networks, more effective knowledge sharing, as well as more effective and efficient task performance. Scott Poole's work (e.g., Poole & Van de Ven, 2004; Van de Ven & Poole, 1995) on organizational change is also relevant in this context, especially the discussion of the four theories of life cycle, teleological, dialectical, and evolutionary change. Furthermore, Contractor's older (e.g., Contractor, 1998; Contractor & Seibold, 1993) and more contemporary work (e.g., Lungeanu, Cater, DeChurch, & Contractor, 2018; Wax, DeChurch, & Contractor, 2017) on team assembly and self-assembly (or self-organizing) grounded in systems theory also speaks to group structuration.

This tradition should be continued and enhanced by computational research, which should not shy away from it. Much inductive work has been spent on reading new forms of organization in online groups under the rubric of "social network community detection." For example, Fortunato (2010) describes a network of scientists and papers by linking them together by authorship. The communities emerge as many linkages join nodes of the same cluster, and fewer linkages join nodes in different clusters. He further argued that the linkages can also be weighted by the number of coauthored papers, to describe communities in a sophisticated way. Similarly, a full typology for aggregative mechanisms is still a work in progress. A huge majority of the studies conducted within the computational framework assumes a homophily effect and preferential time attachment (Jin, Girvan, & Newman, 2001; Mislove, Koppula, Gummadi, Druschel, & Bhattacharjee, 2008; Wagner & Leydesdorff, 2005).

Similar to the notion of homophily is the concept of assortativity (Newman, 2002), which refers to a node's tendency to be connected to other similar nodes in the network (Buccafurri, Lax, & Nocera, 2015). While a majority of the literature supports this claim, Hu and Wang (2009) noted the possibility of disassortative or neutral mixing in online social networks, such as in the case of the platform *Wealink*. Whether referring to homophily or assortativity, the central claim is that similar nodes are always supposed to attach themselves to similar nodes and that the earlier nodes will get more connected in time. Older and newer work, including some empirical research (Brunswicker, Bertino, & Matei, 2015), advises that a variety of models should be considered, including those in which attachment can be directed by exchange of complementarity needs (Monge & Contractor, 2003). Future work should blend these theoretical concerns with the empirical research.

Another fruitful area for investigation is and should be the actual outcome and productivity of online groups. Patrashkova and McComb (2004) performed a computer simulation of cross-functional teams and found that there is a curvilinear relationship between communication and team performance. They explained that too little communication, either synchronous or asynchronous, is clearly bad for team performance. However, too much communication will overload team members with irrelevant information, thus leading to a diminishing return. The arguments intersect with recent findings of computational research on Stack Overflow, which show a curvilinear relationship between content quality and effort evenness across the entire online social organization (Matei, Jabal, & Bertino, 2018).

In a study of online computer games, Gandolfi (2018) found that cooperative thinking and computational thinking are positively correlated. Computational thinking was conceptualized as having the four dimensions of decomposition, abstraction, debugging, and generalization. Gandolfi (2018, p. 132) measured these dimensions with survey items such as "You are able to read a match and its dynamics" (decomposition), "You use game strategies and 'lessons' beyond X (e.g., other games, everyday life) (abstraction), "You constantly update your game strategies" (debugging), and "You are able to adapt yourself in dealing with other situations (e.g., different maps and different players)" (generalization). The finding has implications for the increasing need to promote computational thinking within the education, organizations, communities, and our society. As the online world is created by the massive amount of social interactions by Internet users, the increase in computational thinking will also positively promote online cooperation toward public goods. In fact, this study speaks to Budhathoki and Haythornhwaite's (2012) investigation of motivation for open collaboration on wiki. Budhathoki and Haythornhwaite found that serious and casual contributors differ in their motivations for online contribution toward building an open-source application called OpenStreetMap. They reported that serious mappers were more interested in community and learning as well as local knowledge and career motivations, while casual mappers were more driven by the general principles of making mapping data free and available for the world.

Computational organizational research with a communicative dimension cannot be limited to social media or online investigations. Communication data can come from a variety of sources, including interpersonal exchanges mediated by low-tech or no-tech means. These can be captured by emails (Klimt & Yang, 2004), process records such as meetings, orders, responses, and inquiries (Van Der Aalst, Reijers, & Song, 2005; Van der Aalst & Song, 2004), or by observational research using video records (Cristani, Raghavendra, Del Bue, & Murino, 2013; Yu, Lim, Patwardhan, & Krahnstoever, 2009). Cross sensitizes us to the need not to focus on high-tech mediated communication when trying to understand organizational communication and networks (Cross, Borgatti, & Parker, 2002). He observes that informal networks need to be tracked and contextualized as much as any other types of networks. Thus, computational social network analysis applied to communication cannot and should not be left to the online realm, but continuously expanded in the wake of the latest communication and social network analysis research.

2.4.2 | Trust, credibility, and transparency

Computational communication research is to a large extent social media research. While it does occasionally look at traditional media and other types of textual communication (covered in the content analysis section), its focus has been directed at how social media content is produced and reproduced. An important characteristic of social media is that it produces most of its content through the distributed efforts of its members, which are "prosumers," that is, content producers and consumers. The radical diffusion of production encouraged by social media raises the twin issues of trust and credibility. If anyone can be a producer of information and if any unit of content can be produced by anyone, how can we know that anything is credible?

The crisis of online credibility has started early, with the emergence of the reputation and review systems that score and evaluate online information or the products or people behind them. Sophisticated work has been conducted by computer scientists on automatic the process of crediblization or vetting of content. Research on Amazon reviews (e.g., Liang, 2016; Walther, Liang, Ganster, Wohn, & Emington, 2012), on Slashdot (e.g., Gosain, 2003; Stevenson, 2016) and Reddit (e.g., Kilgo et al., 2016; Lim, Carman, & Wong, 2017; Park, Conway, & Chen, 2018), as well as on social media likes, shares, and comments (e.g., Alhabash, Almutairi, Lou, & Kim, 2018; Liang & Kee, 2018), have opened the road to a very productive line of research. The core research agenda item is discovering methods to identify and mitigate or eliminate false or malevolent content or evaluations. Although brought to the forefront by the recent controversy regarding Donald Trump's election, during which internal and external propaganda machines attempted to influence the outcome, the sources of the problem are much older and significant.

Computational communication scientists working with computer scientists, psychologists, sociologists, political scientists, experts in human–computer interaction, statisticians, and engineers have proposed new approaches to roles, trust, reputation, and transparency in social media (Matei et al., 2015) propose a trailblazing effort for the common research in this area (Matei & Bertino, 2015). Shneiderman's (2015) work, in this context, is crucial. His review of research and recommendation for future work focus on user trust, credible content, reliable sources, and responsible organizations. For each focal area, Shneiderman recommends a specific approach and research agenda. User trust can be measured and gained via network analysis and network interventions. Content credibility can be measured via formal and substantive measures, including evaluative and standard sourcing measures.

In this area, it is notable that the work conducted by computer scientists does not yet intersect with the theoretical work on online credibility heuristics. Projects like TweetCred (http://twitdigest.iiitd.edu.in/TweetCred/), which focuses on formal markers of credibility—semantic, emotional, source provenance—would benefit from theoretical insights such as those suggested by Metzger. In a series of articles (Metzger & Flanagin, 2013; Metzger, Flanagin, & Medders, 2010), Metzger et al. proposed a methodology for ascertaining credibility utilizing which includes sociopsychological processes, such as expectancy violation, personal reputation veracity, and endorsement evidence. The dialogue between computer scientists and communication scholars is particularly necessary, as the crisis of confidence in social media brought about by the 2016 elections showed (Kollanyi, Howard, & Wolley, 2016). Many believe that social media companies and their software engineers showed an abundance of confidence in machine learning and automated vetting methods in checking and eliminating malevolent content generated by automatic and manual foreign agents. Although Twitter claims (Twitter, 2018) that a relatively small percentage of Hillary Clinton's and Donald Trump's were affected by foreign amplification or vilification methods, the challenge of determining what is credible or not online is one of the most important ones that computational communication research needs to deal with.

2.4.3 | Data journalism

Social media and automated methods of data collection create new opportunities for professional journalists, as well. Journalists are story creators, and stories are always about people. To be significant, stories should provide generalizable meanings that speak to individuals. They should explain how and why large groups of individuals did or are about to do certain things, how the choices of some impact others, or how institutions impact everyday lives. Stories rely on anecdotes, but these life slices are often partial and subject to the vagaries of human subjectivity and selective memory. Twitter or Facebook feeds, which can be turned into tabular data by API connections, data scrapping from web pages or automated downloads of Wikipedia edits represent a new motherlode of stories for journalism. Powerful visualizations and analytic processes have been instrumented under the rubric of data journalism. Yet, much more remains to be determined, including the boundaries that need to be put around data practices that can be considered true data journalism.

Data journalism has one root in an older set of practices, computer-assisted journalism (Coddington, 2015). More recently, data journalism has incorporated automatic visualization and data sorting. As with other types of data practices discussed in this article, there is a constant clash between blending human deductive methods and machine derived inductive discovery. The challenge is even more significant in this area, as theory- or story-driven narratives are under the constant threat of automation due to economic and institutional pressures.

Regardless of these tensions, there is significant agreement that data journalism represents a true opportunity. Gynnild (2014) argues that the future of journalism is driven by the news professionals' orientation toward innovation and their ability to think abstractly and work collaboratively with multidisciplinary experts, data, and technologies in the context of journalism. Parasie and Dagiral (2012) further point out the emergence of a group of "programmer–journalists."

Some of the data journalism projects that raise to the level of moments of discovery in a scholarly sense include the powerful visualization of how prevalent and real was the impact of infectious diseases before the introduction of widespread, state-mandated vaccinations (DeBold, 2015). A Peruvian project, "Candidates and Millions" (Amancio, 2016) that tracks contributions to political campaigns have successfully overcome the self-imposed limitations of visualizations, aiming to explore the causal factors (human and institutional) that lead to political corruption. The Pew Internet visualization of polarization (Dimock, Kiley, Keeter, & Doherty, 2014), which preceded the Trump administration, is equally powerful, showing how pivotal moments, especially the dramatic political shift during the 2004–2010 period, have spurred on the political polarization of the nation. While between 2004 and 2008, the Republican voters became more conservative, paradoxically, the democratic voters widened the gap while in power by expressing more liberal opinions.

During the 2016 campaign, a deep data learning project looked at Donald Trump's Twitter metadata, identifying two hardware signatures for the content, one for an iPhone and the other for an Android phone. An analysis of the texts released through the two phones revealed significant stylistic and thematic differences (Robinson, 2016) between the messages issued from the two devices. While the iPhone tweets were more issue and campaign oriented, the Android ones were more emotionally charged, using often attack words and insults (e.g., "dumb" or "crazy"). More importantly, the Android tweets were more negative than those issued on the iPhone, showing a real split in the campaign, in which the staffers are at times working at cross purposes with their boss. Interestingly, another analysis (Matei, 2016), of the more current words used in the Trump and Clinton campaigns showed that both had as their most frequent term "Trump," which put Hillary Clinton at a real disadvantage since she played mostly defense. What connects this type of textual analysis is the fact that data did not quite "speak" on its own. Neither study was completely "inductive." The communicative differences within the Trump campaign, as well as the differences across the two campaigns, were found in a quest to identify, based on a priori understanding of the significant distances between the content creators (Trump, Clinton, staff members).

More recently, Woolley and Howard (2016) highlighted in more theoretical terms the dark side of computational production. They speak about "computational propaganda," defined as "the assemblage of social media platforms, autonomous agents, and big data tasked with the manipulation of public opinion" (p. 4886). Therefore, it is important to understand the algorithms behind these autonomous agents, how they produce political propaganda, and to carefully consider the ethical implications of such algorithmic agents.

This type of research raises the question, again, of what mission and what types of stories will the journalism of the future generate. This will probably be a much more grounded and broader type of analytic discourse, which will combine social insights gained by traditional, qualitative means combined with the powerful harvesting and collation in explicative models of heterogeneous data.

2.5 | Methods of computational communication research

While the various studies reviewed in the previous section of domains review the diversity of methods employed in computational communication research, we highlight two of the most commonly employed methods of content analysis and inferential network analysis. As mentioned earlier, many studies in computational social science treat "text-as-data" (Shah et al., 2015). Therefore, we will give content analysis a closer look.

2.5.1 | Content analysis

Computational methods have much to offer to those trying to classify and analyze content. Lewis et al. (2013) in their comprehensive review of computational content analytics approaches enumerate no less than one dozen studies of Twitter that used automated content analysis. However, they also point to the difficulties encountered by most projects, which can be summarized as one simple phrase: natural language processing. Most common content analysis is often far more complex than natural language processing. Natural language processing has made a lot of progress in the last decade, but it still rooted in the semantic paradigm. The focus is on identifying the semantic meaning of certain utterances.

Content analysis is much more than semantic analysis. It is a form of rhetorical analysis; its goal is to get at the unspoken assumptions and values embedded in text. One of the core tasks of content analysis is to detect and untangle polysemy and contextual modifiers. This might require completely new approaches, which should borrow from speech act theory, ontological semantics, and sociolinguistics. Intention analysis (Cui & Mao, 2017; Purohit, Dong, Shalin, Thirunarayan, & Sheth, 2015) is also a very promising field, which combines some of the simplest methods of detecting the modality with the message (affirmation, negation, exclamation, interrogation) with temporal, sentiment, and topic detection (Curth, 2018).

Sherin (2013) suggested that to a way to resolve the tension between the computational and manual approaches is to employ computational approaches that verify findings derived from the manual and qualitative approach, not simply to replace manual analysis and reduce labor. He argued that the purpose of computational methods is to serve as an independent way to confirm previous analyses and test existing theories. In other words, both should be performed to reinforce each other. In fact, the longest lasting and most challenging aspect of automating content analysis is that content is a moving target, evolving with the theoretical concerns of each study. Content analysis taps into the higher level conceptual interpretation of human discourse, and it will be probably not completely automated until we will have a self-reflexive learning algorithm that can formulate a research problem like a human.

A notable contribution to this effort is the mixed, semantic network analysis proposed by Himelboim, McCreery, and Smith (2013). In their pathbreaking study of controversial political topics on Twitter, they discovered distinct topic and individual clusters finding that Twitter users are not informed by cross-ideological content generated by the other users that they followed. Distinct patterns of use were found for specific political orientations, while blogs and traditional media seemed to interact in a differential manner when shaping political opinions. The study suggested that blending social networks with thematic analysis could be a productive way to analyze social media content.

For researchers interested in content analyzing social media messages, some technical solutions include CONTEXT (Diesner, 2014), AUTOMAP DMI-TCAT (Borra & Rieder, 2014) or smappPy (Social Media and Political Participation Lab at New York University, 2016). Moreover, the eXtensible Text Analysis Suite infrastructure (De Rooij, Vishneuski, & De Rijke, 2012) and the Leipzig Corpus Miner (Niekler, Wiedemann, & Heyer, 2014) are additional options to consider. For more information about scaling up content analysis on big data, see Trilling and Jonkman (2018).

2.5.2 | Predictive network analysis

Earlier in the article, we argued for an emphasis on networked phenomena. Computational research was in fact identified early on as a branch social network analysis (Lazer et al., 2009). Emerging concurrently with the broader field of network science (Barabási, 2003), a good portion of computational research in communication had the ambitious goal to define, characterize, and explain how individuals and groups come together, collaborate on, dispute about, or transact in various outcomes. The network approach was considered vital for describing the social encounters and communication patterns of the new communication environment of the early 21st century. Soon, however, it became quite obvious that describing networks, sometimes with powerful and vibrant visualizations, does not automatically lead to explanatory insights. Most early network analysis was descriptive and inductive. Its main intellectual products were taxonomies of network topologies. While interesting to look at, they many times begged the question: what are you looking at? Fortunately, there has been some breakthrough research in this direction. Smith et al. (2014) provided a new methodology to investigate and categorize Twitter sharing and interaction patterns. Utilizing social network analysis and the versatile NodeXL software platform they have identified six discrete and generalizable network patterns that may explain much of social media message dissemination.

Furthermore, network topologies change over time. This invites the question of what factors determine and condition such transformations. At the forefront of network analytic work in computational communication research, there is a new-fangled science of predictive network analysis, in which either network structures are used to predict node characteristics or node attributes are used to predict the existence or emergence of network edge (ties). Supported by statistical procedures such as exponential random graph modeling (Robins, Pattison, Kalish, & Lusher, 2007) or derivatives of network autocorrelation procedures (Diesner, Frantz, & Carley, 2005; Fujimoto, Chou, & Valente, 2011; Leenders, 2002), network inferential work promises to change the face of computational communication and computational social research. Researchers working in this space can now ask question related to the impact of network ties upon future behavior, the relationship between social capital, communication, and activism, the relationship between social collaboration and public recognition, a fuller method for predicting innovation and adoption of innovation (Kee, Sparks, Struppa, Mannucci, & Damiano, 2016).

The work done in this vein is vast and the contributions too many to be summarized in any meaningful way in an article like this. However, we can mention the research conducted by Contractor et al., Christakis, Williams, Brunswicker, Britt, or

Valente (Britt, 2013; Brunswicker et al., 2015; Christakis & Fowler, 2009; Contractor, DeChurch, Carson, Carter, & Keegan, 2012; Valente, 2010; Williams, Contractor, Poole, Srivastava, & Cai, 2011).

The keys to the success of this work are technical and theoretical. Technically speaking, network analytics of inferential kind demand complex and large-scale calculations, which increase exponentially with the size of the network. This demands nontrivial amounts of computing, both in terms of CPU cycles and RAM. Yet, the more vital need is to specify strong predictive models, which take advantage of the latest advances in predictive network analysis. Review articles, such as Cranmer, Leifeld, McClurg, and Rolfe (2017) or in a more specialized context or by Contractor, Wasserman, and Faust (2006) outline the main methods and possibilities in this arena.

3 | FUTURE WORK

This article aimed to define and exemplify the main areas of interest in computational communication research. The review focused on the tension between inductive and theory-driven approaches, stressing the need to combine the two. Throughout the entire article, we emphasize the gaps left by purely inductive research. Computational models cannot be purely inductive. Computational models will always be as good as the questions that they set out to explore, which come with some prior ideas, formal or informal. These are often driven by inferential mechanisms, which demands attention to chains of partial or complete effect, to antecedent factors and their connection to outcomes.

Furthermore, computational communication research will reach maturity only when content and natural language processing methods will become sophisticated enough to get at the “rhetorical,” that is, persuasive, intentions of the message. Communication research demands and uses content analysis routinely. However, content analysis is more than semantic analysis. It requires coding mechanisms that can capture higher level meanings, intentions, and cultural content.

In a similar vein, computational communication research needs new ways to explain human communicative behaviors. This cannot be reduced to behaviorism, to stimuli and responses. Complex sociopsychological theories have been developed by communication researchers, which include cognitive and sociological factors as independent variables. Recovering and blending this older literature with the newest methodologies is of high importance.

Computational communication research can, at the same time, carve out an individuality of its own by creating big data workflows that can tally, articulate in working models, and use for predictive outcomes a variety of data flows, some of which go beyond social media. This new science of communication will continue and even crown the true vocation of communication research, which is interdisciplinarity.

Computational communication research needs to also propose methods to bootstrap new theories or quasi-theories using partial insights and data. A good amount of social scientific work in communication relies on a combination of methods that resemble trained and untrained machine learning. We have “untrained” methods, such as cluster, factor, or principal component analysis and “trained” hierarchical or simple regression model fitting. What we do not have yet, are more complex, neural network procedures which move rather fast from data to generalizable patterns of behavior that open-up novel theoretical propositions. Although neural network procedures in some respect resemble structural equation modeling, we should not ignore the fact that a good amount of computational research has advanced to the stage where the “latent factors” (neural layers) are now directly inferable from the data, without much supervision. Can these methods be applied with any degree of external validity to social and communication research? Can we, in other words, engage in theory building from the ground up? This is a desideratum that is in much need to clarification and if possible solving.

4 | CONCLUSIONS

New domains of inquiry generate new methods. New methods bring with them new questions, which goad the research community to expand into new domains. The circle of knowledge never stops growing. At each rotation on the spiral of knowledge, some preconceptions are discarded, some new ideas are gained, and many more new challenges are discovered. Computational social and communication science have already brought to the forefront new ways to understand the production of knowledge, diffusion of information, the structuration of large groups, and automated discovery of some simple latent messages. There are, however, limits to what our current computing automata, trained to be only as good as their programs, can do. The next generation of inquiry methods in social science and communication should strive to blend more closely meaning discovery, theory building, and causal inference. These might not be tractable by mere computational clustering or even “neural layer” inference. The polysemy of human language is truly bewildering because language is intrinsically actional. Communication is a tool of human action and interaction. Understanding human meanings needs to situate the knowing subject in the embodied universe of human experience. In a way, the next revolution in communication research is to define and develop a system that not only thinks but also "feels" human. This might be quite a long way, away. Until then, the computational approach will be for sure perfected at least in scale and accessibility. Much remains to be done in terms of bringing it closer to non-computer scientists. Similarly, the conclusions of this research need to be better integrated into future tools and methods of discovery. Thus, the vocation of computational research remains an open-ended one.

CONFLICT OF INTEREST

The authors have declared no conflicts of interest for this article.

RELATED WIREs ARTICLE

Social cognition

REFERENCES

"Foot in Mouth" Prize for Rumsfeld. (2003). Retrieved from http://www.cnn.com/2003/WORLD/europe/12/01/rumsfeld.english.reut/

Abraham, E. P. (1980). Fleming's discovery. *Reviews of Infectious Diseases*, *2*(1), 140–140. https://doi.org/10.1093/clinids/2.1.140

Alhabash, S., Almutairi, N., Lou, C., & Kim, W. (2018). Pathways to virality: Psychophysiological responses preceding likes, shares, comments, and status updates on Facebook. *Media Psychology*, *22*(2), 1–21. https://doi.org/10.1080/15213269.2017.1416296

Amancio, N. l. (2016). Candidates and millions|community. Retrieved from http://community.globaleditorsnetwork.org/node/23093

Anderson, C. (2008). The end of theory: The data deluge makes the scientific method obsolete. *Science* https://www.wired.com/2008/06/pb-theory/

Bacon, F. (2000). *The new organon*. Cambridge, England: Cambridge University Press.

Bakshy, E., Rosenn, I., Marlow, C., & Adamic, L. (2012). The role of social networks in information diffusion. In *Proceedings of the 21st international conference on World Wide Web* (pp. 519–528). Lyon, France: ACM.

Ball-Rokeach, S. J. (1985). The origins of individual media system dependency: A sociological framework. *Communication Research*, *12*, 485–510.

Barabási, A.-L. (2003). *Linked: How everything is connected to everything else and what it means for business, science, and everyday life*. New York: Plume. Retrieved from http://www.amazon.com/gp/product/0452284392?ie=UTF8&tag=paginicom-20&linkCode=as2&camp=1789&creative=390957&creativeASIN=0452284392

Behar, R. (2011, November 25). *FeBI's "Phoenix" memo unmasked [News]*. Retrieved from https://web.archive.org/web/20111125235247/http://features.blogs.fortune.cnn.com/2011/10/12/fbi-phoenix-memo/

Birkinshaw, J., & Ridderstråle, J. (2010). Adhocracy for an agile age. *Organization Science*, *22*(5), 1286–1296.

Blumer, H. (1933). *Movies, delinquency, and crime*. New York, NY: Arno Press.

Borge-Holthoefer, J., Moreno, Y., & Yasseri, T. (2016). At the crossroads: Lessons and challenges in computational social science. *Frontiers in Physics*, *4*, 37.

Borra, E., & Rieder, B. (2014). Programmed method: Developing a toolset for capturing and analyzing tweets. *Aslib Journal of Information Management*, *66*, 262–278. https://doi.org/10.1108/AJIM-09-2013-0094

Brafman, O., & Beckstrom, R. (2006). *The starfish and the spider: The unstoppable power of leaderless organizations*. New York, NY: Portfolio.

Briscoe, B., Odlyzko, A., & Tilly, B. (2006). Metcalfe's law is wrong—Communications networks increase in value as they add members-but by how much? *IEEE Spectrum*, *43*(7), 34–39. https://doi.org/10.1109/MSPEC.2006.1653003

Britt, B. C. (2013). Evolution and revolution of organizational configurations on Wikipedia: A longitudinal network analysis. (Theses and dissertations). ProQuest, 1–164.

Brunswicker, S., Bertino, E., & Matei, S. (2015). Big data for open digital innovation—A research roadmap. *Big Data Research*, *2*, 53–58. https://doi.org/10.1016/j.bdr.2015.01.008

Bryant, J., Thompson, S., & Finklea, B. W. (2012). *Fundamentals of media effects* (2nd ed.). Long Grove, IL: Waveland Pr Inc.

Buccafurri, F., Lax, G., & Nocera, A. (2015). A new form of assortativity in online social networks. *International Journal of Human-Computer Studies*, *80*, 56–65. https://doi.org/10.1016/j.ijhcs.2015.03.006

Budhathoki, N. R., & Haythornhwaite, C. (2012). Motivation for open collaboration: Crowd and community model and the case of OpenStreetMap. *American Behavioral Scientist*, *57*, 548–575.

Burke, M., Marlow, C., & Lento, T. (2009). Feed me: Motivating newcomer contribution in social network sites. In *Proceedings of the SIGCHI conference on human factors in computing systems* (pp. 945–954). New York, NY, USA: ACM. https://doi.org/10.1145/1518701.1518847

Cappella, J. N. (2017). Vectors into the future of mass and interpersonal communication: Big data, social media, and computational social sciences. *Human Communication Research*, *43*, 545–558. https://doi.org/10.1111/hcre.12114

De Choudhury, M., Counts, S., & Horvitz, E. (2013). Social media as a measurement tool of depression in populations. In *Proceedings of the 5th annual ACM Web science conference* (pp. 47–56). Paris, France: ACM.

Christakis, N. A., & Fowler, J. H. (2009). *Connected: The surprising power of our social networks and how they shape our lives*. Columbus, Georgia: Little, Brown and Company.

Churches, O., Nicholls, M., Thiessen, M., Kohler, M., & Keage, H. (2014). Emoticons in mind: An event-related potential study. *Social Neuroscience*, *9*, 196–202. https://doi.org/10.1080/17470919.2013.873737

Coddington, M. (2015). Clarifying journalism's quantitative turn. *Digital Journalism*, *3*, 331–348.

Contractor, N. S. (1998). Self-organizing systems research in the social sciences: Reconciling the metaphors and the models. *Management Communication Quarterly*, *13*, 154–166.

Contractor, N. S., DeChurch, L. A., Carson, J., Carter, D. R., & Keegan, B. (2012). The topology of collective leadership. *The Leadership Quarterly*, *23*(6), 994–1011.

Contractor, N. S., & Seibold, D. R. (1993). Theoretical frameworks for the study of structuring processes in group decision support systems: Adaptive structuration theory and self-organizing systems theory. *Human Communication Research*, *19*, 528–563.

Contractor, N. S., Wasserman, S., & Faust, K. (2006). Testing multitheoretical, multilevel hypotheses about organizational networks: An analytic framework and empirical example. *Academy of Management Review*, *31*(3), 681–703. https://doi.org/10.5465/amr.2006.21318925

Conway, M., & O'Connor, D. (2016). Social media, big data, and mental health: Current advances and ethical implications. *Current Opinion in Psychology*, *9*, 77–82. https://doi.org/10.1016/j.copsyc.2016.01.004

Cranmer, S. J., Leifeld, P., McClurg, S. D., & Rolfe, M. (2017). Navigating the range of statistical tools for inferential network analysis. *American Journal of Political Science*, *61*(1), 237–251. https://doi.org/10.1111/ajps.12263

Cristani, M., Raghavendra, R., Del Bue, A., & Murino, V. (2013). Human behavior analysis in video surveillance: A social signal processing perspective. *Neurocomputing*, *100*, 86–97.

Cross, R., Borgatti, S. P., & Parker, A. (2002). Making invisible work visible: Using social network analysis to support strategic collaboration. *California Management Review*, *44*(2), 25–46.

Crowston, K. (2017). Levels of trace data for social and behavioural science research. In N. Jullien, S. A. Matei, & S. Goggins (Eds.), *Big data factories* (pp. 39–49). New York: Springer.

Crowston, K., Jullien, N., & Ortega, F. (2011, November 16). Too few new wikipedians? Modelling effort and participation in wikipedia. Modelling Effort and Participation in Wikipedia. Retrieved from http://papers.ssrn.com/sol3/papers.cfm?abstract_id=1960696

Crowston, K., Jullien, N., & Ortega, F. (2013). *Is Wikipedia inefficient? Modelling effort and participation in Wikipedia.* Paper presented at the System Sciences (HICSS), 2013 46th Hawaii International Conference on System Sciences.

Cui, C., & Mao, W. (2017). Recognize user intents in online interactions from massive social media data. In *In 2017 I.E. 2nd international conference on big data analysis (ICBDA)* (pp. 11–15). Beijing, China: IEEE. https://doi.org/10.1109/ICBDA.2017.8078805

Curth, J. (2018). Rhetorical Republic: A lexically driven taxonomy for political campaign interactions on Twitter (Master thesis). West Lafayette, IN: Purdue University.

De Rooij, O., Vishneuski, A., & De Rijke, M. (2012). xTAS: Text analysis in a timely manner. In *Proceedings of the 12th Dutch-Belgian information retrieval workshop (DIR 2012)* (pp. 89–90). Twente, Netherlands: University of Twente.

DeBold, T. (2015). Battling infectious diseases in the 20th century: The impact of vaccines. *Wall Street Journal*. Retrieved from http://graphics.wsj.com/infectious-diseases-and-vaccines

Diesner, J. (2014). *ConText: Software for the integrated analysis of text data and network data.* Paper presented at the social and semantic networks in communication research. Preconference at conference of international communication association (ICA), Seattle, WA.

Diesner, J., Frantz, T. L., & Carley, K. M. (2005). Communication networks from the Enron email corpus "It's always about the people. Enron is no different.". *Computational & Mathematical Organization Theory*, *11*(3), 201–228.

Dimock, M., Kiley, J., Keeter, S., & Doherty, C. (2014, June 12). Political Polarization in the American Public. Retrieved from http://www.people-press.org/2014/06/12/political-polarization-in-the-american-public/

Dolan, T. E. (2010). Revisiting adhocracy: From rhetorical revisionism to smart mobs. *Journal of Futures Studies*, *15*(2), 33–50.

Dow, P. A., Adamic, L. A., & Friggeri, A. (2013). The anatomy of large Facebook cascades. *ICWSM*, *1*(2), 12.

Fortunato, S. (2010). Community detection in graphs. *Physics Reports*, *486*, 75–174. https://doi.org/10.1016/j.physrep.2009.11.002

Freelon, D. (2014). On the interpretation of digital trace data in communication and social computing research. *Journal of Broadcast & Electronic Media*, *58*, 59–75. https://doi.org/10.1080/08838151.2013.875018

Fujimoto, K., Chou, C.-P., & Valente, T. W. (2011). The network autocorrelation model using two-mode data: Affiliation exposure and potential bias in the autocorrelation parameter. *Social Networks*, *33*(3), 231–243.

Gandolfi, E. (2018). You have got a (different) friend in me: Asymmetrical roles in gaming as potential ambassadors of computational and cooperative thinking. *E-Learning and Digital Media*, *15*, 128–145.

García-Peñalvo, F. J. (2016). What computational thinking is. *Journal of Information Technology Research*, *9*, v–viii.

Girard, J., & Girard, J. A. (2009). *A Leader's guide to knowledge management: Drawing on the past to enhance future performance*. New York: Business Expert Press.

Gonzalez-Bailon, S., Wang, N., Rivero, A., Borge-Holthoefer, J., & Moreno, Y. (2014). Assessing the bias in samples of large online networks. *Social Networks*, *38*, 16–27. https://doi.org/10.1016/j.socnet.2014.01.004

Gosain, S. (2003). Looking through a window on open source culture: Lessons for community infrastructure design. *Systèmes d'Information et Management*, *8*(1), 1 http://aisel.aisnet.org/sim/vol8/iss1/2

Gynnild, A. (2014). Journalism innovation leads to innovation journalism: The impact of computational exploration on changing mindsets. *Journalism*, *15*, 713–730.

Hagen, J. B. (2000). The origins of bioinformatics. *Nature Reviews. Genetics*, *1*(3), 231–236. https://doi.org/10.1038/35042090

Hansen, D. L., Schneiderman, B., & Smith, M. A. (2010). *Analyzing social media networks with NodeXL insights from a connected world*. Amsterdam; Boston: M. Kaufmann. Retrieved from http://www.sciencedirect.com/science/book/9780123822291

Himelboim, I., McCreery, S., & Smith, M. (2013). Birds of a feather tweet together: Integrating network and content analyses to examine Cross-ideology exposure on twitter. *Journal of Computer-Mediated Communication*, *18*(2), 154–174. https://doi.org/10.1111/jcc4.12001

Hollingshead, A. B., Fulk, J., & Monge, P. (2002). Fostering intranet knowledge sharing: An integration of transactive memory and public goods approaches. In P. J. Hinds & S. Kiesler (Eds.), *Distributed work* (pp. 335–355). Cambridge, MA: MIT Press.

Hu, H.-B., & Wang, X.-F. (2009). Disassortative mixing in online social networks. *EPL: Europhysics Letters*, *86*, 18003. https://doi.org/10.1209/0295-5075/86/18003

Jin, E. M., Girvan, M., & Newman, M. E. J. (2001). Structure of growing social networks. *Physical Review E*, *64*(4), 046132. https://doi.org/10.1103/PhysRevE.64.046132

Johnson, S. (2001). *Emergence: The connected lives of ants, brains, cities, and software*. New York, NY: Scribner.

Joyce, E., & Kraut, R. E. (2006). Predicting continued participation in newsgroups. *Journal of Computer-Mediated Communication*, *11*(3), 723–747.

Katz, E. (2001). Lazarsfeld'S map of media effects. *International Journal of Public Opinion Research*, *13*(3), 270–279. https://doi.org/10.1093/ijpor/13.3.270

Kee, K. F., Sparks, L., Struppa, D. C., Mannucci, M., & Damiano, A. (2016). Information diffusion, Facebook clusters, and the simplicial model of social aggregation: A computational simulation of simplicial diffusers for community health interventions. *Health Communication*, *31*, 385–399. https://doi.org/10.1080/10410236.2014.960061

Kilgo, D. K., Yoo, J. J., Sinta, V., Geise, S., Suran, M., & Johnson, T. J. (2016). Let it on Reddit: An exploratory study examining opinion leadership on Raddit. *First Monday*. https://doi.org/10.5210/fm.v21i9.6429

Kittur, A., Lee, B., & Kraut, R. E. (2009). Coordination in collective intelligence: The role of team structure and task interdependence. In *Proceedings of the 27th international conference on human factors in computing systems* (pp. 1495–1504). Boston, MA: ACM. https://doi.org/10.1145/1518701.1518928

Kittur, A., Nickerson, J. V., Bernstein, M., Gerber, E., Shaw, A., Zimmerman, J., … Horton, J. (2013). The future of crowd work. In *Proceedings of the 2013 conference on computer supported cooperative work* (pp. 1301–1318). New York, NY: ACM. https://doi.org/10.1145/2441776.2441923

Klimt, B., & Yang, Y. (2004). The enron corpus: A new dataset for email classification research. *European Conference on Machine Learning*, *18*, 217–226. http://doi.org/10.1007/978-3-540-30115-8_22

Kollanyi, B., Howard, P. N., & Wolley, S. C. (2016). *Bots and Automation over Twitter during the U.S. Election (Data Memo)*. Project on Computation Propaganda. Oxford, England: Oxford Internet Institute. Retrieved from https://www.oii.ox.ac.uk/blog/bots-and-automation-over-twitter-during-the-u-s-election/

Konieczny, P. (2010). Adhocratic governance in the internet age: A case of Wikipedia. *Journal of Information Technology & Politics*, *7*(4), 263–283.

Kramer, A. D. I., Guillory, J. E., & Hancock, J. T. (2014). Experimental evidence of massive-scale emotional contagion through social networks. *Proceedings of the National Academy of Sciences*, *111*(24), 8788–8790. https://doi.org/10.1073/pnas.1320040111

Lampe, C., & Johnston, E. (2005). Follow the (slash) dot: Effects of feedback on new members in an online community. In *Proceedings of the 2005 international ACM SIGGROUP conference on Supporting group work* (pp. 11–20). Sanibel Island, FL: ACM.

Lasswell, H. (1968). The function and structure of communication. In L. Bryson (Ed.), *The communication of ideas: A series of addresses* (pp. 37–52). New York, NY: Harper and Brothers. Retrieved from https://books.google.com/books?id=3BoEAAAAMAAJ

Lazer, D., Pentland, A. S., Adamic, L., Aral, S., Barabasi, A.-L., Brewer, D., & Gutmann, M. (2009). Social science: Computational social science. *Science*, *323*, 721–723. https://doi.org/10.1126/science.1167742

Leenders, R. T. A. (2002). Modeling social influence through network autocorrelation: Constructing the weight matrix. *Social Networks*, *24*(1), 21–47.

Leskovec, J., & Sosič, R. (2016). SNAP: A general-purpose network analysis and graph-mining library. *ACM Transactions on Intelligent Systems and Technology (TIST)*, *8*(1), 1–20.

Lewis, S. C., Zamith, R., & Hermida, A. (2013). Content analysis in an era of big data: A hybrid approach to computational and manual methods. *Journal of Broadcasting & Electronic Media*, *57*, 34–52. https://doi.org/10.1080/08838151.2012.761702

Liang, Y. J. (2016). Reading to make a decision or to reduce cognitive dissonance? The effect of selecting and reading online reviews from a post-decision context. *Computers in Human Behavior*, *64*, 463–471. https://doi.org/10.1016/j.chb.2016.07.016

Liang, Y. J., & Kee, K. F. (2018). Developing and validating the ABC framework of information diffusion on social media. *New Media & Society*, *20*, 272–292. https://doi.org/10.1177/1461444816661552

Lim, W. H., Carman, M. J., & Wong, S. J. (2017). Estimating relative user expertise for content quality prediction on Reddit. In *HT '17 Proceedings of the 28th ACM conference on hypertext and social media* (pp. 55–64). Prague, Czech Republic: ACM. https://doi.org/10.1145/3078714.3078720

Logan, D. C. (2009). Known knowns, known unknowns, unknown unknowns and the propagation of scientific enquiry. *Journal of Experimental Botany*, *60*(3), 712–714. https://doi.org/10.1093/jxb/erp043

Luft, J., & Ingham, H. (1955). *The Johari window: A graphic model for interpersonal relations*. Los Angeles: University of California Western Training Lab.

Lungeanu, A., Cater, D. R., DeChurch, L. A., & Contractor, N. (2018). How team interlock ecosystems shape the assembly of scientific teams: A hypergraph approach. *Communication Methods and Measures*, *12*, 174–198. https://doi.org/10.1080/19312458.2018.1430756

Marty, E., Rebillard, F., Pouchot, S., & Lafouge, T. (2012). Diversity and concentration of online news. *Réseaux*, *176*(6), 27–72.

Matei, S. A. (2016, June 23). The message of Trump is Trump. And Hillary Clinton's, too! [Research]. Retrieved from http://matei.org/ithink/2016/06/23/the-message-of-trump-is-trump-and-and-hillary-clintons-too/

Matei, S. A., & Bertino, E. (2015). *Roles, trust, and reputation in social media knowledge markets: Theory and methods*. New York, NY: Springer Nature.

Matei, S. A., & Britt, B. C. (2017). *Structural differentiation in social media: Adhocracy, entropy, and the "1% effect"* (1st ed.). New York, NY: Springer Nature.

Matei, S. A., Jabal, A. A., & Bertino, E. (2017). Do sticky elites produce online knowledge of higher quality? In *Proceedings of the 2017 IEEE/ACM international conference on advances in social networks analysis and mining 2017* (pp. 72–79). New York, NY: ACM https://doi.org/10.1145/3110025.3110040

Matei, S. A., Jabal, A. A., & Bertino, E. (2018). Social-collaborative determinants of content quality in online knowledge production systems: Comparing Wikipedia and Stack Overflow. *Social Network Analysis and Mining*, *8*(36), 8–36. https://doi.org/10.1007/s13278-018-0512-3

Matei, S. A., Jullien, N., & Goggins, S. P. (Eds.). (2017). *Big data factories: Collaborative approaches*. New York: Springer International Publishing. Retrieved from //www.springer.com/us/book/9783319591858

Matei, S. A., Russell, M. G., & Bertino, E. (Eds.). (2015). *Transparency in social media*. New York, NY: Springer Nature.

Merton, R. K., Fiske, M., & Curtis, A. (1946). *Mass persuasion: The social psychology of a war bond drive*. New York, NY: Harper & Bros.

Metzger, M. J., & Flanagin, A. J. (2013). Credibility and trust of information in online environments: The use of cognitive heuristics. *Journal of Pragmatics*, *59*, 210–220.

Metzger, M. J., Flanagin, A. J., & Medders, R. B. (2010). Social and heuristic approaches to credibility evaluation online. *Journal of Communication*, *60*(3), 413–439.

Mislove, A., Koppula, H. S., Gummadi, K. P., Druschel, P., & Bhattacharjee, B. (2008). Growth of the Flickr social network. In *Proceedings of the first workshop on online social networks* (pp. 25–30). New York, NY: ACM. https://doi.org/10.1145/1397735.1397742

Monge, P. R., & Contractor, N. (2003). *Theories of communication networks* (1st ed.). Oxford; New York: Oxford University Press.

Nesvizhskii, A. I. (2014). Proteogenomics: Concepts, applications and computational strategies. *Nature Methods*, *11*(11), 1114–1125. https://doi.org/10.1038/nmeth.3144

Newman, M. (2010). *Networks: An introduction* (1st ed.). Oxford; New York: Oxford University Press.

Newman, M. E. J. (2002). Assortative mixing in networks. *arXiv.org* https://arxiv.org/abs/cond-mat/0205405

Niekler, A., Wiedemann, G., & Heyer, G. (2014). Leipzig Corpus Miner—A text mining infrastructure for qualitative data analysis. In *Terminology and knowledge engineering (TKE)*. Berlin: Informatics Institute, University of Leipzig.

O'Donnell, M. B., & Falk, E. B. (2015). Big data under the microscope and brains in social context: Integrating methods from computational social science and neuroscience. *The Annals of the American Academy of Political and Social Science*, *659*(1), 274–289.

Ortega, F., Gonzalez-Barahona, J. M., & Robles, G. (2008). On the inequality of contributions to Wikipedia. In *Proceedings of the 41st annual Hawaii international conference on system sciences* (p. 304). Waikoloa, HI: IEEE. https://doi.org/10.1109/HICSS.2008.333

Parasie, S., & Dagiral, E. (2012). Data-driven journalism and the public good: "Computer-assisted-reporters" and "programmer-journalists" in Chicago. *New Media & Society*, *15*, 853–871.

Park, A., Conway, M., & Chen, A. T. (2018). Examining thematic similarity, difference, and membership in three online mental health communities from Reddit: A text mining and visualization approach. *Computers in Human Behavior*, *78*, 98–112. https://doi.org/10.1016/j.chb.2017.09.001

Park, S., Shim, H. S., Chatterjee, M., Sagae, K., & Morency, L.-P. (2014). Computational analysis of persuasiveness in social multimedia: A novel dataset and multimodal prediction approach. In *Proceedings of the 16th ACM international conference on multimodal interaction (ICMI '14)*. Istanbul, Turkey: ACM.

Patrashkova, R. R., & McComb, S. A. (2004). Exploring why more communication is not better: Insights from a computational model of cross-functional teams. *Journal of Engineering and Technology Management*, *21*(1–2), 83–114. https://doi.org/10.1016/j.jengtecman.2003.12.005

Pont, M. T. S., Mora, H. M., Casado, G. D. M., & Mendez, D. G. (2016). A computational model of the belief system under the scope of social communication. *Foundation of Science*, *21*, 215–223.

Poole, M. S., & Van de Ven, A. H. (2004). *Handbook of organizational change and innovation*. Oxford, UK: Oxford University Press.

Purohit, H., Dong, G., Shalin, V., Thirunarayan, K., & Sheth, A. (2015). Intent classification of short-text on social media. In *2015 I.E. international conference on Smart City/SocialCom/SustainCom (SmartCity)* (pp. 222–228). Chengdu, China: IEEE.

Rebillard, F., & Smyrnaios, N. (2010). Les infomédiaires, au cø eur de la filière de l'information en ligne. *Réseaux*, *160-161*(2), 163–194.

Robins, G., Pattison, P., Kalish, Y., & Lusher, D. (2007). An introduction to exponential random graph (p*) models for social networks. *Social Networks*, *29*, 173–191. https://doi.org/10.1016/j.socnet.2006.08.002

Robinson, D. (2016, August 9). Text analysis of Trump's tweets confirms he writes only the (angrier) Android half [Blog]. Retrieved from http://varianceexplained.org/r/trump-tweets/

Rumsfeld, D. (2002, February 12). Defense.gov Transcript: DoD News Briefing—Secretary Rumsfeld and Gen. Myers. Retrieved from http://archive.defense.gov/Transcripts/Transcript.aspx?TranscriptID=2636

Shah, D. V., Cappella, J. N., & Neuman, W. R. (2015). *Big data, digital media, and computational social science: Possibilities and perils* (Vol. 659). ANNALS, AAPSS. https://doi.org/10.1177/0002716215572084

Shao, W., Ross, M., & Grace, D. (2015). Developing a motivation-based segmentation typology of Facebook users. *Marketing Intelligence & Planning*, *33*(7), 1071–1086. https://doi.org/10.1108/MIP-01-2014-0014

Shaw, A., & Hill, B. M. (2014). Laboratories of Oligarchy? How the iron law extends to peer production. *Journal of Communication*, *64*(2), 215–238. https://doi.org/10.1111/jcom.12082

Sherin, B. (2013). A computational study of commonsense science: An exploration in the automated analysis of clinical interview data. *Journal of the Learning Sciences*, *22*(4), 600–638. https://doi.org/10.1080/10508406.2013.836654

Sherman, L. E., Payton, A. A., Hernandez, L. M., Greenfield, P. M., & Dapretto, M. (2016). The power of the like in adolescence: Effects of peer influence on neural and behavioral responses to social media. *Psychological Science*, *27*, 1027–1035. https://doi.org/10.1177/0956797616645673

Shneiderman, B. (2015). Building trusted social media communities: A research roadmap for promoting credible content. In *Roles, trust, and reputation in social media knowledge markets* (pp. 35–43). Cham: Springer https://doi.org/10.1007/978-3-319-05467-4_2

Smith, A., & Anderson, M. (2018, March 1). Social Media Use in 2018. Retrieved from http://www.pewinternet.org/2018/03/01/social-media-use-in-2018/.

Smith, Marc, Shneiderman, Ben, & Himelboim, I. (2014, February 20). Mapping twitter topic networks: From polarized crowds to community clusters. Retrieved from http://www.pewinternet.org/2014/02/20/mapping-twitter-topic-networks-from-polarized-crowds-to-community-clusters/.

Social Media and Political Participation Lab at New York University. (2016). smapPy. Retrieved from https://github.com/SMAPPNYU/smappPy

Sparks, G. G. (2015). *Media effects research: A basic overview* (5th ed.). Australia: Wadsworth Publishing.

Stevenson, M. (2016). Slashdot, open news and informated media: exploring the intersection of imagined futures and web publishing technology. In W. H. K. Chun, A. W. Fisher & T. Keenan (Eds.), *New Media, Old Media: A History and Theory Reader* (pp. 616–630). London: Routledge.

Surowiecki, J. (2004). *The wisdom of crowds: Why the many are smarter than the few and how collective wisdom shapes business, economies, societies, and nations* (1st ed.). New York, NY: Doubleday.

Trilling, D., & Jonkman, J. G. F. (2018). Scaling up content analysis. *Communication Methods and Measures*, *12*, 158–174. https://doi.org/10.1080/19312458.2018.1447655

Twitter (2018). Update on results of retrospectice review of Russia-related election activity. In *United States senate on the Judicriary, subcommitte on crime and Terrorism*. Washington, DC: US Senate. Retrieved from https://www.judiciary.senate.gov/imo/media/doc/Edgett%20Appendix%20to%20Responses.pdf

Valente, T. W. (2010). *Social networks and health: Models, methods, and applications* (1st ed.). Oxford: Oxford University Press.

van Atteveldt, W., & Peng, T. (2018). When communication meets computation: Opportunities, challenges, and pitfalls in computational communication research. *Communication Methods and Measures*, *12*, 81–92. https://doi.org/10.1080/19312458.2018.1458084

Van de Ven, A. H., & Poole, M. S. (1995). Explaining development and change in organizations. *Academy of Management Review*, *20*, 510–540.

Van Der Aalst, W. M., Reijers, H. A., & Song, M. (2005). Discovering social networks from event logs. *Computer Supported Cooperative Work (CSCW)*, *14*(6), 549–593.

Van der Aalst, W. M., & Song, M. (2004). Mining social networks: Uncovering interaction patterns in business processes. In *International conference on business process management* (pp. 244–260). New York: Springer.

Wagner, C. S., & Leydesdorff, L. (2005). Network structure, self-organization, and the growth of international collaboration in science. *Research Policy*, *34*(10), 1608–1618. https://doi.org/10.1016/j.respol.2005.08.002

Walther, J. B., Liang, Y., Ganster, T., Wohn, D. Y., & Emington, J. (2012). Online reviews, helpfulness ratings, and consumer attitudes: An extension of congruity theory to multiple sources in Web 2.0. *Journal of Computer-Mediated Communication*, *18*, 97–112.

Wasko, M. M., & Faraj, S. (2000). "It is what one does": Why people participate and help others in electronic communities of practice. *The Journal of Strategic Information Systems*, *9*(2–3), 155–173.

Watts, D. J. (2003). *Six degrees: The science of a connected age* (1st ed.). New York, NY: Norton.

Watts, D. J. (January 2017). Should social science be more solution-oriented? *Nature Human Behaviour*, *1*(1), 0015. https://doi.org/10.1038/s41562-016-0015

Wax, A., DeChurch, L., & Contractor, N. (2017). Self-organizing into winning teams: Understanding the mechanisms that drive successful collaborations. *Small Group Research*, *48*, 665–718. https://doi.org/10.1177/1046496417724209

Welser, H. T., Gleave, E., Fisher, D., & Smith, M. (2007). Visualizing the signatures of social roles in online discussion groups. *Journal of Social Structure*, *8*(2), 1–32.

Welser, H. T., Cosley, D., Kossinets, G., Lin, A., Dokshin, F., Gay, G., & Smith, M. (2011). Finding social roles in Wikipedia. In *Proceedings of the 2011 iConference* (pp. 122–129). Seattle, WA: ACM. Retrieved from http://dl.acm.org/citation.cfm?id=1940778

Williams, D., Contractor, N., Poole, M. S., Srivastava, J., & Cai, D. (2011). The virtual worlds exploratorium: Using large-scale data and computational techniques for communication research. *Communication Methods and Measures*, *5*(2), 163–180.

Wing, J. M. (2006). Computational thinking. *Communications of the ACM*, *49*, 33–35. https://doi.org/10.1145/1118178.1118215

Woolley, S. C., & Howard, P. N. (2016). Political communication, computational propaganda, and autonomous agents. *International Journal of Communication*, *10*, 4882–4890.

Yu, T., Lim, S.-N., Patwardhan, K., & Krahnstoever, N. (2009). Monitoring, recognizing, and discovering social networks. In *IEEE conference on computer vision and pattern recognition. CVPR 2009* (pp. 1462–1469). Miami Beach, FL: IEEE.

Zagorski, N. (2017). Using many social media platforms linked with depression. *Anxiety Risk. Psychiatrics News*. https://doi.org/10.1176/appi.pn.2017.1b16

How to cite this article: Matei SA, Kee KF. Computational communication research. *WIREs Data Mining Knowl Discov*. 2019;e1304. https://doi.org/10.1002/widm.1304